August 30, 2024

VIA ECF

Hon. Ona T. Wang
United States District Court
Southern District of New York

Re: *The New York Times Company v. Microsoft Corporation et al.*,
Case No. 23-cv-11195-SHS-OTW | Dispute Re: NYT Search Terms

Dear Magistrate Judge Wang:

The New York Times ("The Times") again burdens the Court with an unnecessary and premature motion. This time, The Times seeks to compel OpenAI to review half a million documents (beyond documents OpenAI is already reviewing) at an estimated expense of over $[redacted]. The Times' demands are overbroad, unduly burdensome, and the product of The Times' refusal to negotiate in good faith. The Court should deny The Times' motion.

I. The Times' motion is premature, again

The Times has declared an impasse where none exists. The parties last met and conferred on search terms at the end of July. On that call, OpenAI asked The Times to map each of its proposed terms onto specific Requests for Production ("RFPs") because, given the overbreadth and vagueness of The Times' proposals, OpenAI had been unable to determine each term's relevance in order to propose better terms. OpenAI reiterated its request two weeks later. The Times remained silent. Suddenly, after hours on August 28, The Times declared an impasse. It filed its motion 32 minutes later. *See* Dkt. Nos. 228, 228-2. This outright refusal to explain its terms' relevance, coupled with The Times' haste to file, illustrate The Times' refusal to confer in good faith. The Times' motion should be denied on that ground alone. *See* Dkt. Nos. 135-1, at 3 ("The parties will negotiate in good faith . . . to identify custodians and search terms."), 135-2, at 4 (same); J. Wang Indiv. Practices in Civ. Cases II.b.

II. OpenAI has appropriately rejected The Times' search terms as irrelevant, overbroad, unduly burdensome, and disproportionate to the needs of the case

OpenAI is best positioned to determine how to search its own documents. "Absent agreement," unless a producing party's choice about how to conduct a search is "manifestly unreasonable or the requesting party demonstrates that the resulting production is deficient, the court should play no role in dictating the design of the search, whether in choosing search tools [or] selecting search terms[.]" *Mortg. Resol. Servicing, LLC v. JPMorgan Chase Bank, N.A.*, No. 15-cv-0293-LTS-JCF, 2017 WL 2305398, at *2 (S.D.N.Y. May 18, 2017). Here, the Times seeks to dictate the design and scope of OpenAI's search. The Times does not come close to showing OpenAI's choices are "manifestly unreasonable," or any deficient production.

A. The Times has failed to show each term's relevance

Via its motion, The Times has—for the first time—disclosed categories indicating the types of information its terms are *intended* to capture. Still, however, The Times has not mapped those

requests onto valid RFPs. *See* Fed. R. Civ. Proc. 37(a)(3)(B)(iv); *see also Munguia-Brown v. Equity Residential*, 337 F.R.D. 509, 518 (N.D. Cal. 2021) (denying motion to compel in part because requested documents were not responsive to an RFP); *see also* S.D.N.Y L. Civ. R. 37.1 (requiring verbatim discovery requests and responses to which the motion is addressed). This is both inconsistent with The Times' own approach to the discovery it seeks, *see* Ex. B (The Times rejecting terms to apply to its documents as "not directed at any discovery request propounded by OpenAI that The Times agreed to produce documents in response to"), and fatal to The Times' motion. *See* Ex. A (Dkt. No. 228-1 annotated to include OpenAI's objections).

B. The Times has failed to tailor its terms

It is The Times' burden to narrowly tailor its discovery requests. *See Park Miller, LLC v. Durham Grp., Ltd.*, No. 19-cv-04185-WHO, 2020 WL 6047236, at *4 (N.D. Cal. Oct. 13, 2020) (rejecting Plaintiffs' requests as "far from narrowly tailored"). But here, The Times only claims that "each term will capture documents that OpenAI has already agreed to produce." Dkt. 228 at 1. That is only accurate on the most facile level because The Times' imprecise and overbroad terms would *also* sweep in hundreds of thousands of additional documents, including those of little or no relevance. *See* Ex. A.

For instance, The Times argues that term 12 relates to "GenAI and Journalism." *See* Dkt. 228-1, row 19. Yet term 12 does not include the word "journalism" at all. *Id.* And term 14 (including, e.g., "importan*" within 50 of "*gpt*") is allegedly related to "The Times' Legal Claims." But such a broad search hits on more than 145,000 documents, *see* Dkt. 228-1, row 23, including many that will have no bearing on the issues in dispute. Even though The Times' terms may capture some relevant documents, relevance alone is not the standard; discovery must be both relevant *and* "proportional to the needs of the case." Fed. R. Civ. P. 26(b)(1).

Meanwhile, The Times objects to OpenAI's inclusion (or in Plaintiff's language, "recycling") of terms relating to the alleged use of books in OpenAI's training data. Dkt. No. 228 at 1. But The Times propounded numerous discovery requests concerning OpenAI's training data. Ex. C (Excerpts of The Times' First RFPs at Nos. 3, 14); Ex. D (Excerpts of The Times' Second RFPs at Nos. 17, 33, 36). Indeed, The Times acknowledged in the Rule 26(f) Report that its RFPs "substantially overlap with those previously served" in the cases relating to OpenAI's alleged training on books. Dkt. No. 72 at 15-16. The Court should disregard The Times' complaints about OpenAI offering to search for documents responsive to The Times' own RFPs.

C. Search terms are an inappropriate method for providing some information

Now that OpenAI has been provided some information about the types of documents The Times seeks, it is clear that other discovery methods are better suited to providing the requested information. For example, The Times argues that "Terms 59-60 are designed to capture documents concerning licensing agreements." Dkt. No. 228 at 3. These terms are unnecessary because OpenAI had *already* agreed to, and is *already* producing final, executed data access agreements from non-custodial sources.

As another example, The Times proposes terms relating to, among other things, OpenAI's training data, Retrieval Augmented Generation (or "RAG"), and OpenAI's model cards. OpenAI

has, in fact, *already* agreed to the terms "Retrieval Augmented Generation" and "RAG." Regardless, much of The Times' requested information is best provided either through inspection of OpenAI's training data or source code, or through targeted collections. OpenAI has also *already* agreed to produce) its responsive model cards after a targeted collection. The Times' proposal seeks burdensome and inappropriate search terms on these and similar topics. *See* Ex. A.

Here again, The Times applies a different standard to OpenAI than itself. For its own document review, The Times objected to many terms OpenAI proposed because The Times was "producing responsive documents on this topic through targeted, non-custodial collections." *See, e.g.*, Ex. B, cells D25, D27. If The Times is using targeted collections when appropriate, OpenAI should be able to do so as well.

D. The Times' proposed search terms are unduly burdensome

The Times "bears the initial burden of demonstrating that the information sought is relevant and proportional." *See* S*portvision, Inc. v. MLB Advanced Media, L.P.*, No. 18-cv-03025-PGG-VF, 2022 WL 2817141, at *1 (S.D.N.Y. July 19, 2022). Yet The Times demands OpenAI review approximately half a million documents *in addition to* documents OpenAI already reviewed or is reviewing. Dkt. No. 228 at 1, 2; *see* Ex. E. This number is likely an undercount, as it only reflects term hits for OpenAI's initial list of custodians. In a separate premature motion, The Times moved to compel production from 17 additional custodians, Dkt. No. 205, and OpenAI has responded by offering 14, Dkt. No. 212. While those custodians aren't at issue here, if The Times' request were applied to these custodians, it would result in over 1.3 *million* documents to review. Ex. E.

Regardless, whether the review is of 500,000 or 1,300,000 documents, the volume is burdensome and unreasonable. OpenAI's e-discovery vendor estimates these terms would add [redacted] hours of contract review time and $[redacted] in expense. Ex. E. The Times has not justified this enormous demand. S*ee* Fed. R. Civ. P. 26(b)(1), (2)(C) (requiring courts to "limit the frequency or extent of discovery" when "the burden or expense of the proposed discovery outweighs its likely benefit"); *In re Weatherford Int'l Sec. Litig.*, No. 11-cv-01646-LAK-JCF, 2013 WL 2355451, at *5 (S.D.N.Y. May 28, 2013) ("A proportionality analysis requires the court to balance the value of the requested discovery against the cost of its production."). Moreover, The Times *agrees* that undue burden should limit production, at least for its own documents. The Times has objected to reviewing OpenAI's proposed search terms that hit on, for example, 3,057 documents, claiming the terms "are overbroad and directed at too many irrelevant documents." *Compare* Ex. B, cell J38, *with* Dkt. 228 at 2 (The Times arguing "roughly 5,000" OpenAI documents is a "paltry number"). Overall, The Times has agreed to review *at most* 211,695 documents related to search terms.[1] Requiring OpenAI to review half a million (or 1.3 million) documents is simply not proportional.

* * *

Considering the above, OpenAI requests that the Court deny The Times' motion to compel.

[1] The Times has not provided a report of total unique documents. OpenAI suspects the number is much lower.

Dated: August 30, 2024

Respectfully Submitted,

/s/ *Vera Ranieri*
Vera Ranieri (*pro hac vice*)
Morrison & Foerster LLP

/s/ *Elana Nightengale Dawson*
Elana Nightengale Dawson (*pro hac vice*)
Latham & Watkins LLP

/s/ *Michelle Ybarra*
Michelle Ybarra (*pro hac vice*)
Keker, Van Nest & Peters LLP

cc: All Counsel of Record (via ECF)